Don't PANIC: Prototypical Additive Neural Network for Interpretable Classification of Alzheimer's Disease

Tom Nuno Wolf[1,2(✉)], Sebastian Pölsterl[2], and Christian Wachinger[1,2]

[1] Department of Radiology, Technical University Munich, Munich, Germany
tom_nuno.wolf@tum.de
[2] Lab for Artificial Intelligence in Medical Imaging, Ludwig Maximilians University, Munich, Germany

Abstract. Alzheimer's disease (AD) has a complex and multifactorial etiology, which requires integrating information about neuroanatomy, genetics, and cerebrospinal fluid biomarkers for accurate diagnosis. Hence, recent deep learning approaches combined image and tabular information to improve diagnostic performance. However, the black-box nature of such neural networks is still a barrier for clinical applications, in which understanding the decision of a heterogeneous model is integral. We propose PANIC, a prototypical additive neural network for interpretable AD classification that integrates 3D image and tabular data. It is interpretable by design and, thus, avoids the need for post-hoc explanations that try to approximate the decision of a network. Our results demonstrate that PANIC achieves state-of-the-art performance in AD classification, while directly providing local and global explanations. Finally, we show that PANIC extracts biologically meaningful signatures of AD, and satisfies a set of desirable desiderata for trustworthy machine learning. Our implementation is available at https://github.com/ai-med/PANIC.

1 Introduction

It is estimated that the number of people suffering from dementia worldwide will reach 152.8 million by 2050, with Alzheimer's disease (AD) accounting for approximately 60–80% of all cases [20]. Due to large studies, like the Alzheimer's Disease Neuroimaging Initiative (ADNI; [9]), and advances in deep learning, the disease stage of AD can now be predicted relatively accurate [28]. In particular, models utilizing both tabular and image data have shown performance superior to unimodal models [4,5,29]. However, they are considered black-box models, as their decision-making process remains largely opaque. Explaining decisions of Convolutional Neural Networks (CNN) is typically achieved with post-hoc techniques in the form of saliency maps. However, recent studies showed that different post-hoc techniques lead to vastly different explanations of the same

---

T. N. Wolf and S. Pölsterl—These authors contributed equally to this work.

model [12]. Hence, post-hoc methods do not mimic the true model accurately and have low fidelity [22]. Another drawback of post-hoc techniques is that they provide local interpretability only, i.e., an approximation of the decision of a model for a specific input sample, which cannot explain the overall decision-making of a model. Rudin [22] advocated to overcome these shortcomings with inherently interpretable models, which are interpretable by design. For instance, a logistic regression model is inherently interpretable, because one can infer the decision-making process from the weights of each feature. Moreover, inherently interpretable models do provide both local and global explanations. While there has been progress towards inherently interpretable unimodal deep neural networks (DNNs) [1,11], there is a lack of inherently interpretable heterogeneous DNNs that incorporate both 3D image and tabular data.

In this work, we propose PANIC, a Prototypical Additive Neural Network for Interpretable Classification of AD, that is based on the Generalized Additive Model (GAM). PANIC consists of one neural net for 3D image data, one neural net for each tabular feature, and combines their outputs via summation to yield the final prediction (see Fig. 1). PANIC processes 3D images with an inherently interpretable CNN, as proposed in [11]. The CNN is a similarity-based classifier that reasons by comparing latent features of an input image to a set of class-representative latent features. The latter are representations of specific images from the training data. Thus, its decision-making can be considered similar to the way humans reason. Finally, we show that PANIC is fully transparent, because it is interpretable both locally and globally, and achieves state-of-the-art performance for AD classification.

2   Related Work

Interpretable Models for Tabular Data. Decision trees and linear models, such as logistic regression, are inherently interpretable and have been applied widely [16]. In contrast, multi-layer perceptrons (MLPs) are non-parametric and non-linear, but rely on post-hoc techniques for explanations. A GAM is a non-linear model that is fully interpretable, as its prediction is the sum of the outputs of univariate functions (one for each feature) [15]. Explainable Boosting Machines (EBMs) extend GAMs by allowing pairwise interaction of features. While this may boost performance compared to a standard GAM, the model is harder to interpret, because the number of functions to consider grows quadratically with the number of features. EBMs were used in [24] to predict conversion to AD.

Interpretable Models for Medical Images. ProtoPNet [2] is a case-based interpretable CNN that learns class-specific prototypes and defines the prediction as the weighted sum of the similarities of features, extracted from a given input image, to the learned prototypes. It has been applied in the medical domain for diabetic retinopathy grading [6]. One drawback is that prototypes are restricted by the size of local patches: For example, it cannot learn a *single* prototype to represent hippocampal atrophy, because the hippocampus appears in the left and
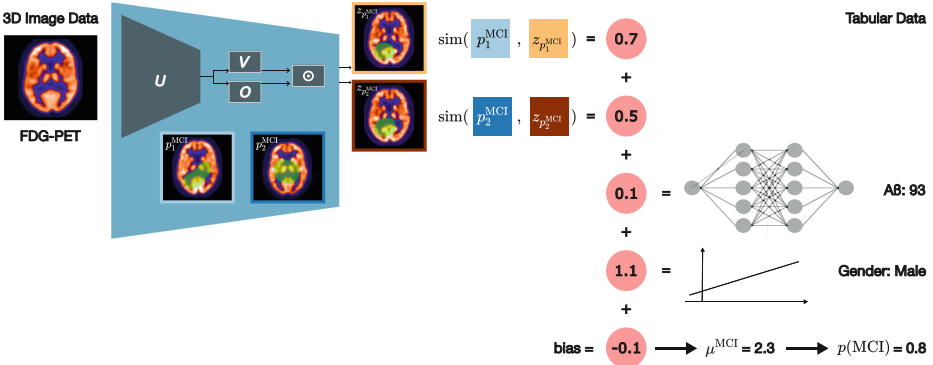
Fig. 1. An exemplary prediction for class MCI with PANIC: 3D FDG-PET images are processed with an interpretable CNN that computes the cosine similarity between latent representations $z_{p_1^{\mathrm{MCI}}}, z_{p_2^{\mathrm{MCI}}}$ and corresponding prototypes $p_1^{\mathrm{MCI}}, p_2^{\mathrm{MCI}}$, as seen on the left. For each categorical feature, such as gender, a linear function is learned. Each continuous feature, such as Aβ is processed with its own MLP. The final prediction is the sum of the outputs of the submodules plus a bias term.

right hemisphere. As a result, the number of prototypes needs to be increased to learn a separate prototype for each hemisphere. The Deformable ProtoPNet [3] allows for multiple fine-grained prototypical parts to extract prototypes, but is bound to a fixed number of prototypical parts that represent a prototype. XProtoNet [11] overcomes this limitation by defining prototypes based on attention masks rather than patches; it has been applied for lung disease classification from radiographic images. Wang et al. [27] used knowledge distillation to guide the training of a ProtoPNet for mammogram classification. However, their final prediction is uninterpretable, because it is the average of the prediction of a ProtoPNet and a black-box CNN. The works in [8,19,21,27,31] proposed interpretable models for medical applications, but in contrast to ProtoPNet, XProtoNet and GAMs, they do not guarantee that explanations are faithful to the model prediction [22].

3 Methods

We propose a prototypical additive neural network for interpretable classification (PANIC) that provides unambiguous local and global explanations for tabular and 3D image data. PANIC leverages the transparency of GAMs by adding functions that measure similarities between an input image and a set of class-specific prototypes, that are latent representations of images from the training data [11].

Let the input consist of N tabular features $x_n \in \mathbb{R}$ ($n \in \{1, \ldots, N\}$), and a 3D grayscale image $\mathcal{I} \in \mathbb{R}^{1 \times H \times D \times W}$. PANIC is a GAM comprising N univariate functions f_n to account for tabular features, and an inherently interpretable

CNN g to account for image data [11]. The latter provides interpretability by learning a set of $K \times C$ class-specific prototypes (C classes, K prototypes per class $c \in \{1, \ldots, C\}$). During inference, the model seeks evidence for the presence of prototypical parts in an image, which can be visualized and interpreted in the image domain. Computing the similarities of prototypes to latent features representing the presence of prototypical parts allows to predict the probability of a sample belonging to class c:

$$p(c \,|\, x_1, \ldots, x_N, \mathcal{I}) = \text{softmax}\left(\mu^c\right), \quad \mu^c = \beta_0^c + \textstyle\sum_{n=1}^{N} f_n^c(x_n) + \sum_{k=1}^{K} g_k^c(\mathcal{I}), \ (1)$$

where $\beta_0^c \in \mathbb{R}$ denotes a bias term, $f_n^c(x_n)$ the class-specific output of a neural additive model for feature n, and $g_k^c(\mathcal{I})$ the similarity between the k-th prototype of class c and the corresponding feature extracted from an input image $\mathcal{I}$. We define the functions f_n^c and g_k^c below.

3.1 Modeling Tabular Data

Tabular data often consists of continuous and discrete-valued features, such as age and genetic alterations. Therefore, we model feature-specific functions f_n^c depending on the type of feature n. If it is continuous, we estimate f_n^c non-parametrically using a multi-layer perceptron (MLP), as proposed in [1]. This assures full interpretability while allowing for non-linear processing of each feature n. If feature n is discrete, we estimate f_n^c parametrically using a linear model, in which case f_n^c is a step function, which is fully interpretable too. Moreover, we explicitly account for missing values by learning a class-conditional missing value indicator s_n^c. To summarize, f_n^c is defined as

$$f_n^c(x_n) = \begin{cases} s_n^c, & \text{if } x_n \text{ is missing,} \\ \beta_n^c x_n, \ \text{with } \beta_n^c \in \mathbb{R}, & \text{if } x_n \text{ is categorical} \\ \text{MLP}_n^c(x_n), & \text{otherwise.} \end{cases} \quad (2)$$

Predicting a class with the sum of such univariate functions f_n^c was proposed in [1] as Neural Additive Model (NAM). Following [1], we apply an ℓ_2 penalty on the outputs of $f_n^c(x_n)$:

$$\mathcal{L}_{\text{Tab}}(x_1, \ldots, x_n) = \tfrac{1}{C} \sum_{c=1}^{C} \sum_{n=1}^{N} [f_n^c(x_n)]^2.$$

We want to emphasize that NAMs retain global interpretability by plotting each univariate function f_n^c over its domain (see Fig. 2). Local interpretability is achieved by evaluating $f_n^c(x_n)$, which equals the contribution of a feature x_n to the prediction of a sample, as defined in Eq. (1) (see Fig. 3 on the left).

3.2 Modeling Image Data

We model 3D image data by defining the function $g_k^c(\mathcal{I})$ in Eq. (1) based on XProtoNet [11], which learns prototypes that can span multiple, disconnected
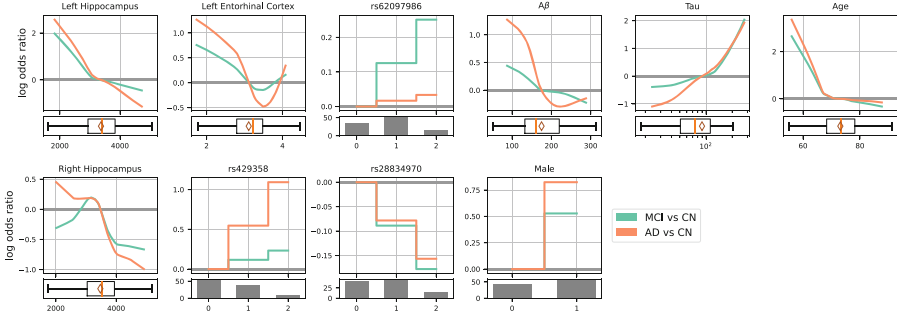
Fig. 2. The plots show the log odds ratio with respect to controls for the top 10 tabular features. Boxplots show the distribution in our data.
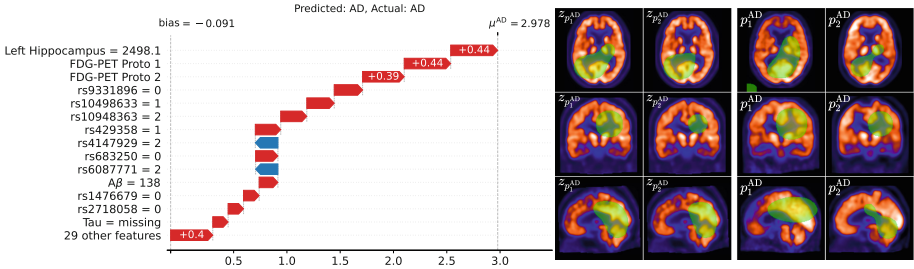

Fig. 3. Explanation for the prediction of a single sample from the test set. Left: Individual contributions to the overall prediction. Right: Input FDG-PET overlayed with the attention maps (green) of the corresponding representations $z_{p_1^{\mathrm{AD}}}, z_{p_1^{\mathrm{AD}}}$ (columns 1,2), and the attention maps of learned prototypes $p_1^{\mathrm{AD}}, p_2^{\mathrm{AD}}$ (columns 3,4). (Color figure online)

regions within an image. In XProtoNet, an image is classified based on the cosine similarity between a latent feature vector $z_{p_k^c}$ and learned class-specific prototypes p_k^c, as depicted in the top part of Fig. 1:

$$g_k^c(\mathcal{I}) = \mathrm{sim}(p_k^c, z_{p_k^c}) = \frac{p_k^c \cdot z_{p_k^c}}{\|p_k^c\| \|z_{p_k^c}\|}. \tag{3}$$

A latent feature vector $z_{p_k^c}$ is obtained by passing an image $\mathcal{I}$ into a CNN backbone $\mathcal{U} : \mathbb{R}^{1 \times H \times D \times W} \rightarrow \mathbb{R}^{R \times H' \times D' \times W'}$, where R is the number of output channels. The result is passed into two separate modules: (i) the feature extractor $\mathcal{V} : \mathbb{R}^{R \times H' \times D' \times W'} \rightarrow \mathbb{R}^{L \times H' \times D' \times W'}$ maps the feature map to the dimensionality of the prototype space L; (ii) the occurrence module $\mathcal{O}^c : \mathbb{R}^{R \times H' \times D' \times W'} \rightarrow \mathbb{R}^{K \times H' \times D' \times W'}$ produces K class-specific attention masks. Finally, the latent feature vector $z_{p_k^c}$ is defined as

$$z_{p_k^c} = \mathrm{GAP}[\mathrm{sigmoid}(\mathcal{O}^c(\mathcal{U}(\mathcal{I}))_k) \odot \mathrm{softplus}(\mathcal{V}(\mathcal{U}(\mathcal{I})))], \tag{4}$$

where $\odot$ denotes the Hadamard product, and GAP global average pooling.

Intuitively, $z_{p_k^c}$ represents the GAP-pooled activation maps that a prototype p_k^c would yield if it were present in that image. For visualization, we can upsample the occurrence map $\mathcal{O}^c(\mathcal{U}(\mathcal{I}))_k$ to the input dimensions and overlay it on the input image. The same can be done to visualize prototype p_k^c (see Fig. 3).

Regularization. Training XProtoNet requires regularization with respect to the occurrence module and prototype space [11]: An occurrence and affine loss enforce sparsity and spatial fidelity of the attention masks $\mathcal{O}^c$ with respect to the image domain:

$$\mathcal{L}_{\text{occ}}(\mathcal{I}) = \sum_{c=1}^{C}\|\mathcal{O}^c(\mathcal{U}(\mathcal{I}))\|_1, \quad \mathcal{L}_{\text{affine}}(\mathcal{I}) = \|A(\mathcal{O}^c(\mathcal{U}(\mathcal{I}))) - \mathcal{O}^c(\mathcal{U}(A(\mathcal{I})))\|_1,$$

with A a random affine transformation. Additionally, latent vectors $z_{p_k^c}$ of an image $\mathcal{I}$ with true class label y should be close to prototypes of their respective class, and distant to prototypes of other classes:

$$\mathcal{L}_{\text{clst}}(\mathcal{I}) = -\max_{k,c=y}\ g_k^c(\mathcal{I}), \quad \mathcal{L}_{\text{sep}}(\mathcal{I}) = \max_{k,c\neq y}\ g_k^c(\mathcal{I}).$$

3.3   PANIC

As stated in Eq. (1), PANIC is a GAM comprising functions $f_1^c,\ldots,f_N^c$ for tabular data, and functions $g_1^c,\ldots,g_K^c$ for 3D image data (see Eqs. (2) and (3)). Tabular features contribute to the overall prediction in Eq. (1) in terms of the values $f_1^c(x_1),\ldots,f_N^c(x_N)$, while the image contributes in terms of the cosine similarity between the class-specific prototype p_k^c and the latent feature vector $z_{p_k^c}$. By restricting prototypes to contribute to the prediction of a specific class only, we encourage the model to learn discriminative prototypes for each class.

To interpret PANIC locally, we simply consider the outputs of the functions f_n^c, and sum the image-based similarity scores over all prototypes: $\sum_{k=1}^{K} g_k^c(\mathcal{I})$. To interpret the contributions due to the 3D image in detail, we examine the attention map of each prototype, the attention map of the input image, and the similarity score between each prototype and the image (see Fig. 3 on the right). To interpret PANIC globally, we compute the absolute contribution of each function to the per-class logits in Eq. (1), and average it over all samples in the training set, as seen in Fig. 4. In addition, we can directly visualize the function f_n^c learned from the tabular data in terms of the log odds ratio

$$\log\left[\frac{p(c\,|\,x_1,\ldots,x_n,\ldots,x_N,\mathcal{I})}{p(\text{CN}\,|\,x_1,\ldots,x_n,\ldots,x_N,\mathcal{I})}\bigg/\frac{p(c\,|,x_1,\ldots,x_n',\ldots,x_N,\mathcal{I})}{p(\text{CN}\,|\,x_1,\ldots,x_n',\ldots,x_N,\mathcal{I})}\right],$$

where x_n' is the mean value of feature n across all samples for continuous features, and zero for categorical features. As an example, let us consider the AD class. If the log odds ratio for a specific value x_n is positive, it indicates that the odds of being diagnosed as AD, compared to CN, increases. Conversely, if it is negative, the odds of being diagnosed as AD decreases.

Table 1. Statistics for the data used in our experiments.

	CN (N = 379)	Dementia (N = 256)	MCI (N = 610)	Total (N = 1245)
Age				
Mean (SD)	73.5 (5.9)	74.5 (7.9)	72.3 (7.3)	73.1 (7.1)
Range	55.8–90.1	55.1–90.3	55.0–91.4	55.0–91.4
Gender				
Female	193 (50.9%)	104 (40.6%)	253 (41.5%)	550 (44.2%)
Male	186 (49.1%)	152 (59.4%)	357 (58.5%)	695 (55.8%)
Education				
Mean (SD)	16.4 (2.7)	15.4 (2.8)	16.1 (2.7)	16.0 (2.8)
Range	7.0–20.0	4.0–20.0	7.0–20.0	4.0–20.0
MMSE				
Mean (SD)	29.0 (1.2)	23.2 (2.2)	27.8 (1.7)	27.2 (2.7)
Range	24.0–30.0	18.0–29.0	23.0–30.0	18.0–30.0

We train PANIC with the following loss:

$$\mathcal{L}(y, x_1, \ldots, x_n, \mathcal{I}) = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \lambda_1 \mathcal{L}_{\text{Tab}}(x_1, \ldots, x_n) + \lambda_2 \mathcal{L}_{\text{clst}}(\mathcal{I})$$
$$+ \lambda_3 \mathcal{L}_{\text{sep}}(\mathcal{I}) + \lambda_4 \mathcal{L}_{\text{occ}}(\mathcal{I}) + \lambda_5 \mathcal{L}_{\text{affine}}(\mathcal{I}), \quad (5)$$
$$\hat{y} = \arg\max_c \; p(c \,|\, x_1, \ldots, x_n, \mathcal{I}),$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss, $\lambda_{1,\ldots,5}$ are hyper-parameters, y the true class label, and $\hat{y}$ the prediction of PANIC.

4   Experiments

4.1   Overview

Dataset. Data used in this work was obtained from the ADNI database.[1] We select only baseline visits to avoid data leakage, and use FDG-PET images following the processing pipeline described in [18]. Tabular data consists of the continuous features age, education; the cerobrospinal fluid markers Aβ, Tau, p-Tau; the MRI-derived volumes of the left/right hippocampus and thickness of the left/right entorhinal cortex. The categorical features are gender and 31 AD-related genetic variants identified in [7,13] and having a minor allele frequency of $\geq 5\%$. Tabular data was standardized using the mean and standard deviation across the training set. Table 1 summarizes our data.

---

[1] https://adni.loni.usc.edu/ [9].

Implementation Details. We train PANIC with the loss in Eq. (5) with AdamW [14] and a cyclic learning rate scheduler [26], with a learning rate of 0.002, and weight decay of 0.0005. We choose a 3D ResNet18 for the CNN backbone $\mathcal{U}$ with $R = 256$ channels in the last ResBlock. The feature extractor $\mathcal{V}$ and the occurrence module $\mathcal{O}$ are CNNs consisting of $1 \times 1 \times 1$ convolutional layers with ReLU activations. We set $K = 2$, $L = 64$, $\lambda_1 = 0.01$ and $\lambda_{2,...,5} = 0.5$, and norm the length of each prototype vector to one. For each continuous tabular feature n, f_n^c is a MLP that shares parameters for the class dependent outputs in (2). Thus, each MLP has 2 layers with 32 neurons, followed by an output layer with C neurons. As opposed to [1], we found it helpful to replace the ExU activations by ReLU activations. We apply spectral normalization [17] to make MLPs Lipschitz continuous. We add dropout with a probability of 0.4 to MLPs, and with a probability of 0.1 to all univariate functions in Eq. (1). The set of affine transformations A comprises all transformations with a scale factor $\in [0.8; 1.2]$ and random rotation $\in [-180°; 180°]$ around the origin. We initialize the weights of $\mathcal{U}$ from a pre-trained model that has been trained on the same training data on the classification task for 100 epochs with early stopping. We cycle between optimizing all parameters of the network and optimizing the parameters of f_n^c only. We only validate the model directly after prototypes p_k^c have been replaced with their closest latent feature vector $z_{p_k^c}$ of samples from the training set of the same class. Otherwise, interpretability on an image level would be lost.

We perform 5-fold cross-validation, based on a data stratification strategy that accounts for sex, age and diagnosis. Each training set is again split such that 64% remain training and 16% are used for hyper-parameter tuning (validation set). We report the mean and standard deviation of the balanced accuracy (BAcc) of the best hyper-parameters found on the validation sets.

4.2   Classification Performance

PANIC achieves 64.0 ± 4.5% validation BAcc and 60.7 ± 4.4% test BAcc. We compare PANIC to a black-box model for heterogeneous data, namely DAFT [29]. We carry out a random hyper-parameter search with 100 configurations for learning rate, weight decay and the bottleneck factor of DAFT. DAFT achieves a validation and test BAcc of 60.9 ± 0.7% and 56.2 ± 4.5%, respectively. This indicates that interpretability does not necessitate a loss in prediction performance.

4.3   Interpretability

PANIC is easy to interpret on a global and local level. Figure 4 summarizes the average feature importance over the training set. It shows that FDG-PET has on average the biggest influence on the prediction, but also that importance can vary across classes. For instance, the SNP rs429358, which is located in the ApoE gene, plays a minor role for the controls, but is highly important for the AD class. This is reassuring, as it is a well known risk factor in AD [25]. The
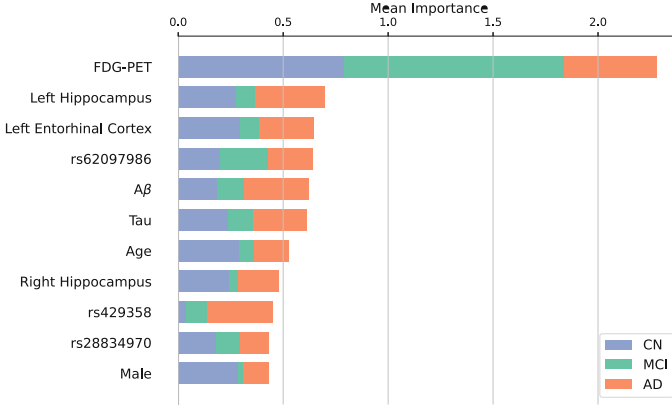
Fig. 4. Ranking of the most import features learned by PANIC. FDG-PET denotes the combined importance of all prototypes related to the FDG-PET image.

overall most important SNP is rs62097986, which is suspected to influence brain volumes early in neurodevelopment [7].

To get a more detailed inside into PANIC, we visualize the log odds ratio with respect to the function f_n^c across the domain of the nine most important tabular features in Fig. 2. We can easily see that PANIC learned that atrophy of the left hippocampus increases the odds of being diagnosed as AD. The volume of the right hippocampus is utilized similarly. For MCI, it appears as PANIC has overfit on outliers with very low right hippocampus volume. Overall, the results for left/right hippocampus agree with the observation that the hippocampus is among the first structures experiencing neurodegeneration in AD [23]. The function for left entorhinal cortex thickness agrees with previous results too [23]. An increase in Aβ measured in CSF is associated with a decreased risk of AD [25], which our model captured correctly. The inverse relationship holds for Tau [25], which PANIC confirmed too. The function learned for age shows a slight decrease in the log odds ratio of AD and MCI, except for patients around 60 years of age, which is due to few data samples for this age range in the training data. We note that the underlying causes that explain the evolution from normal aging to AD remain unknown, but since age is considered a confounder one should control for it [10]. Overall, we observe that PANIC learned a highly non-linear function for the continuous features hippocampus volume, entorhinal thickness, Aβ, Tau, and age, which illustrates that estimating the functions f_n^c via MLPs is effective. In our data, males have a higher incidence of AD (see Table 1), which is reflected in the decision-making of the model too. Our result that rs28834970 decreases the odds for AD does not agree with previous results [13]. However, since PANIC is fully interpretable, we can easily spot this misconception.

Additionally, we can visualize the prototypes by upscaling the attention map specific to each prototype, as produced by the occurrence module, to the input image size and highlighting activations of more than 30% of the maximum value,

as proposed in [11] (see Fig. 3 on the right). The axial view of p_1^{AD} shows attention towards the occipital lobe and p_2^{AD} towards one side of the occipital lobe. Atrophy around the ventricles can be seen in FDG-PET [25] and both prototypes p_1^{AD} and p_2^{AD} incorporate this information, as seen in the coronal views. The sagittal views show, that p_2^{AD} focuses on the cerebellum and parts of the occipital lobe. The parietal lobe is clearly activated by the prototype in the sagittal view of p_1^{AD} and was linked to AD previously [25].

We can interpret the decision-making of PANIC for a specific subject by evaluating the contribution of each function with respect to the prediction (see Fig. 3 on the left). The patient was correctly classified as AD, and most evidence supports this (red arrows). The only exceptions are the SNPs rs4147929 and rs6087771, which the models treats as evidence against AD. Hippocampus volume contributed most to the prediction, i.e., was most important. Since, the subject's left hippocampus volume is relatively low (see Fig. 2), this increases our trust in the model's prediction. The subject is heterozygous for rs429358 (ApoE), a well known marker of AD, which the model captured correctly [25]. The four variants rs9331896, rs10498633, rs4147929, and rs4147929 have been associated with nucleus accumbens volume [7], which is involved in episodic memory function [30]. Atrophy of the nucleus accumbens is associated with cognitive impairment [30]. FDG-PET specific image features present in the image show a similarity of 0.44 to the features of prototype p_1^{AD}. It is followed by minor evidence of features similar to prototype p_2^{AD}. During the prediction of the test subject, the network extracted prototypical parts from similar regions. As seen in the axial view, both parts $z_{p_1^{\mathrm{AD}}}$ and $z_{p_2^{\mathrm{AD}}}$ contain parts of the occipital lobe. Additionally, they cover a large part of the temporal lobe, which has been linked to AD [25].

In summary, the decision-making of PANIC is easy to comprehend and predominantly agrees with current knowledge about AD.

5   Desiderata for Machine Learning Models

We now show that PANIC satisfies four desirable desiderata for machine learning (ML) models, based on the work in [22].

Explanations Must be Faithful to the Underlying Model (Perfect Fidelity). To avoid misconception, an explanation must not mimic the underlying model, but equal the model. The explanations provided by PANIC are the values provided by the functions $f_1^c, \ldots, f_N^c, g_1^c, \ldots, g_K^c$ in Eq. (1). Since PANIC is a GAM, the sum of these values (plus bias) equals the prediction. Hence, explanations of PANIC are faithful to how the model arrived at a specific prediction. We can plot these values to gain local interpretability, as done in Fig. 3.

Explanations Must be Detailed. An explanation must be comprehensive such that it provides all information about what and how information is used by a model (global interpretability). The information learned by PANIC can be described precisely. Since PANIC is based on the sum of univariate functions, we

can inspect individual functions to understand what the model learned. Plotting the functions f_n^c over their domain explains the change in odds when the value x_n changes, as seen in Fig. 2. For image data, PANIC uses the similarity between the features extracted from the input image and the K class-specific prototypes. Global interpretability is achieved by visualizing the training image a prototype was mapped to, and its corresponding attention map 3.

A Machine Learning Model Should Help to Improve the Knowledge Discovery Process. For AD, the precise cause of cognitive decline remains elusive. Therefore, ML should help to identify biomarkers and relationships, and inform researchers studying the biological causes of AD. PANIC is a GAM, which means it provides full global interpretability. Therefore, the insights it learned from data are directly accessible, and provide unambiguous feedback to the knowledge discovery process. For instance, we can directly infer what PANIC learned about FDG-PET or a specific genetic mutation, as in Figs. 2 and 4. This establishes a feedback loop connecting ML researchers and researchers studying the biological causes of AD, which will ultimately make diagnosis more accurate.

ML Models Must be Easy to Troubleshoot. If an ML model produces a wrong or unexpected result, we must be able to easily troubleshoot the model. Since PANIC provides local and global interpretability, we can easily do this. We can use a local explanation (see Fig. 3) to precisely determine what the deciding factor for the prediction was and whether it agrees with our current understanding of AD. If we identified the culprit, we can inspect the function f_n^c, in the case of a tabular feature, or the prototypes, in the case of the image data: Suppose, the age of the patient in Fig. 3 was falsely recorded as 30 instead of 71. The contribution of age $f_{\mathrm{age}}^{\mathrm{AD}}$ would increase from -0.138 to 3.28, thus, dominate the prediction by a large margin. Hence, the local explanation would reveal that something is amiss and prompt us to investigate the learned function $f_{\mathrm{age}}^{\mathrm{AD}}$ (see Fig. 2), which is ill-defined for this age.

6   Conclusion

We proposed an inherently interpretable neural network for tabular and 3D image data, and showcased its use for AD classification. We used local and global interpretability properties of PANIC to verify that the decision-making of our model largely agrees with current knowledge about AD, and is easy to troubleshoot. Our model outperformed a state-of-the-art black-box model and satisfies a set of desirable desiderata that establish trustworthiness in PANIC.

References

1. Agarwal, R., Melnick, L., Frosst, N., et al.: Neural additive models: interpretable machine learning with neural nets. In: NeurIPS, vol. 34, pp. 4699–4711 (2021)
2. Chen, C., Li, O., Tao, D., et al.: This looks like that: deep learning for interpretable image recognition. In: NeurIPS, vol. 32 (2019)
3. Donnelly, J., Barnett, A.J., Chen, C.: Deformable ProtoPNet: an interpretable image classifier using deformable prototypes. In: CVPR, pp. 10265–10275 (2022)
4. El-Sappagh, S., Abuhmed, T., Islam, S.M.R., Kwak, K.S.: Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data. Neurocomputing **412**, 197–215 (2020)
5. Esmaeilzadeh, S., Belivanis, D.I., Pohl, K.M., Adeli, E.: End-To-end Alzheimer's disease diagnosis and biomarker identification. In: Shi, Y., Suk, H.-I., Liu, M. (eds.) MLMI 2018. LNCS, vol. 11046, pp. 337–345. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00919-9_39
6. Hesse, L.S., Namburete, A.I.L.: INSightR-Net: interpretable neural network for regression using similarity-based comparisons to prototypical examples. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 502–511. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_48
7. Hibar, D.P., Stein, J.L., Renteria, M.E., et al.: Common genetic variants influence human subcortical brain structures. Nature **520**(7546), 224–9 (2015)
8. Ilanchezian, I., Kobak, D., Faber, H., Ziemssen, F., Berens, P., Ayhan, M.S.: Interpretable gender classification from retinal fundus images using BagNets. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 477–487. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_45
9. Jack, C.R., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging **27**(4), 685–691 (2008)
10. Jagust, W.: Imaging the evolution and pathophysiology of Alzheimer disease. Nat. Rev. Neurosci. **19**(11), 687–700 (2018)
11. Kim, E., Kim, S., Seo, M., Yoon, S.: XProtoNet: diagnosis in chest radiography with global and local explanations. In: CVPR, pp. 15719–15728 (2021)
12. Kindermans, P.-J., et al.: The (un)reliability of saliency methods. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 267–280. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_14
13. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., et al.: Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. **45**(12), 1452–1458 (2013)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
15. Lou, Y., Caruana, R., Gehrke, J.: Intelligible models for classification and regression. In: SIGKDD, pp. 150–158 (2012)
16. Martino, F.D., Delmastro, F.: Explainable AI for clinical and remote health applications: a survey on tabular and time series data. Artif. Intell. Rev. **56**, 5261–5315 (2022)
17. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018)
18. Narazani, M., Sarasua, I., Pölsterl, S., et al.: Is a PET all you need? a multi-modal study for Alzheimer's disease using 3D CNNs. In: Wang, L., Dou, Q., Fletcher,

P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 66–76. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_7

19. Nguyen, H.D., Clément, M., Mansencal, B., Coupé, P.: Interpretable differential diagnosis for Alzheimer's disease and frontotemporal dementia. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13431, pp. 55–65. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_6

20. Nichols, E., Steinmetz, J.D., Vollset, S.E., et al.: Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. Lancet Public Health **7**(2), e105–e125 (2022)

21. Oba, Y., Tezuka, T., Sanuki, M., Wagatsuma, Y.: Interpretable prediction of diabetes from tabular health screening records using an attentional neural network. In: DSAA, pp. 1–11 (2021)

22. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)

23. Sabuncu, M.R.: The dynamics of cortical and hippocampal atrophy in Alzheimer disease. Arch. Neurol. **68**(8), 1040 (2011)

24. Sarica, A., Quattrone, A., Quattrone, A.: Explainable boosting machine for predicting Alzheimer's disease from MRI hippocampal subfields. In: Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q., Zhong, N. (eds.) BI 2021. LNCS (LNAI), vol. 12960, pp. 341–350. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86993-9_31

25. Scheltens, P., Blennow, K., Breteler, M.M.B., et al.: Alzheimer's disease. Lancet **388**(10043), 505–517 (2016)

26. Smith, L.N.: Cyclical learning rates for training neural networks. In: WACV, pp. 464–472 (2017)

27. Wang, C., Chen, Y., Liu, Y., et al.: Knowledge distillation to ensemble global and interpretable prototype-based mammogram classification models. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022, vol. 13433, pp. 14–24. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_2

28. Wen, J., et al.: Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med. Image Anal. **63**, 101694 (2020)

29. Wolf, T.N., Pölsterl, S., Wachinger, C.: DAFT: a universal module to interweave tabular data and 3D images in CNNs. NeuroImage **260**, 119505 (2022)

30. Yi, H.A., Möller, C., Dieleman, N., et al.: Relation between subcortical grey matter atrophy and conversion from mild cognitive impairment to Alzheimer's disease. J. Neurol. Neurosurg. Psychiatry **87**(4), 425–432 (2015)

31. Yin, C., Liu, S., Shao, R., Yuen, P.C.: Focusing on clinically interpretable features: selective attention regularization for liver biopsy image classification. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 153–162. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_15